Editorial: "Signals to Signs" – Feature Extraction, Recognition, and Multimodal Fusion

Kostas Karpouzis

Abstract Processing of recorded or real-time signals, feature extraction, and recognition are concepts of utmost important to an affect–aware and capable system, since they offer the opportunity to machines to benefit from modeling human behavior based on theory and interpret it based on observation. This chapter discusses feature extraction and recognition based on unimodal features in the case of speech, facial expressions and gestures, and physiological signals and elaborates on attention, fusion, dynamics, and adaptation in different multimodal cases.

Signal processing, feature extraction, and recognition are integral parts of an affect–aware system. The central role of these processes is illustrated in the "map of the thematic areas involved in emotion-oriented computing" included in the "Start here!" section of the HUMAINE portal (Fig. 1)

Here, emotion detection as a whole is strongly connected to "raw empirical data," represented by the "Databases" chapters of this handbook, "usability and evaluation" and "synthesis of basic signs," and also has strong links to "theory of emotional processes." With respect to databases, this Handbook Area discusses the algorithms used to extract features from individual modalities from natural, naturalistic, and acted audiovisual data, the approaches used to provide automatic annotation of unimodal and multimodal data, taking into account different emotion representations as described in the Theory chapters, and the fallback approaches which can be used when the unconstrained nature of these data hampers extraction of detailed features (this also involves usability concepts). In addition to this, studies correlating manual annotation to automatic classification of expressivity have been performed in order to investigate the extent to which the latter can introduce a pre-processing step to the annotation of large audiovisual databases.

Regarding synthesis and embodied conversational agents (ECAs), this chapter discusses how low-level features (e.g., raising eyebrows or hand movements) can

K. Karpouzis (✉)
Image, Video and Multimedia Systems Lab, Institute of Communications and Computer Systems, National Technical University of Athens, Athens, Greece
e-mail: kkarpou@image.ece.ntua.gr

P. Petta et al. (eds.), *Emotion-Oriented Systems*, Cognitive Technologies,
DOI 10.1007/978-3-642-15184-2_5,

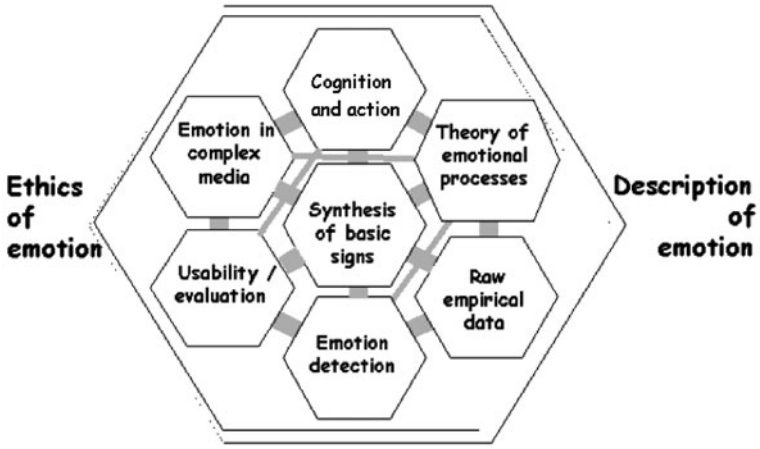

Fig. 1 Map of the thematic areas involved in emotion-oriented computing (from http://emotion-research.net/aboutHUMAINE/start-here, visited 2010-05-31)

be connected to higher level concepts (facial expressions or semantic gestures) using emotion representation theories, how ideas from image processing and scene analysis can be utilized in virtual environments, supplying ECAs with attention capabilities, and how real-time feature extraction from facial expressions and hand gestures can be used to render feedback-capable ECAs.

1 Multimodality and Interaction

The misconceptions related to multimodality in emotion recognition and interaction were discussed in great detail by Oviatt (1999). However, a number of comparative studies (Castellano et al., 2008; Gunes and Piccardi, 2009) illustrate that taking into account multiple channels, either in terms of features or in terms of unimodal decisions or labels, does benefit recognition rates and robustness. Systems can integrate signals at the feature level (Rogozan, 1999) or, after coming up with a class decision at the feature level of each modality, by merging decisions at a semantic level (late identification, Rogozan, 1999; Teissier et al., 1999), possibly taking into account any confidence measures provided by each modality or, generally, a mixture of experts mechanism. Cognitive modeling and experiments indicate that this kind of fusion may happen at feature level (Onat et al., 2007) but discussion regarding the semantics and robustness of each approach is still open in this area.

The inherent multimodality of human interaction can also be exploited in terms of complementing information across channels as well; consider, for instance, that human speech is bimodal in nature (Chen and Rao, 1998). Speech that is perceived by a person depends not only on acoustic cues but also on visual cues such as lip movements or facial expressions. This combination of auditory and visual speech

recognition is more accurate than auditory only or visual only, since use of multiple sources generally enhances speech perception and understanding. Consequently, there has been a large amount of research on incorporating bimodality of speech into human–computer interaction interfaces. Lip sync is one of the research topics in this area (Zoric and Pandzic, 2006).

For interactive applications it is necessary to perform lip sync in real time, which is a particular challenge not only because of computational load but also because the low delay requirement reduces the audio frame available for analysis to a minimum. Speech sound is produced by vibration of the vocal cords and then it is additionally modeled by vocal tract (Lewis and Parke, 1986). A phoneme, defined as basic unit of acoustic speech, is determined by vocal tract, while intonation characteristics (pitch, amplitude, voiced/whispered quality) are dependent on the sound source. Lip synchronization is the determination of the motion of the mouth and tongue during speech (McAllister et al., 1997). To make lip synchronization possible, position of the mouth and tongue must be related to characteristics of the speech signal. Positions of the mouth and tongue are functions of the phoneme and are independent of intonation characteristics of speech.

There are many sounds that are visually ambiguous when pronounced. Therefore, there is a many-to-one mapping between phonemes and visemes, where viseme is a visual representation of phoneme (Pandžić and Forchheimer, 2002).

The process of automatic lip sync, as shown in Fig. 2, consists of two main parts (Zoric, 2003). The first one, audio to visual mapping, is a key issue in bimodal speech processing. In this first phase, speech is analyzed and classified into viseme categories. In the second part, calculated visemes are used for animation of virtual character’s face. Audio to visual (AV) mapping can be solved on several different levels, depending on the speech analysis that is being used. In Zoric (2003), speech is classified into viseme classes by neural networks and GA is used for obtaining the optimal neural network topology. By introducing segmentation of the speech directly into viseme classes instead of phoneme classes, computation overhead is reduced, since only visemes are used for facial animation. Automatic design of neural networks with genetic algorithms saves much time in the training process.

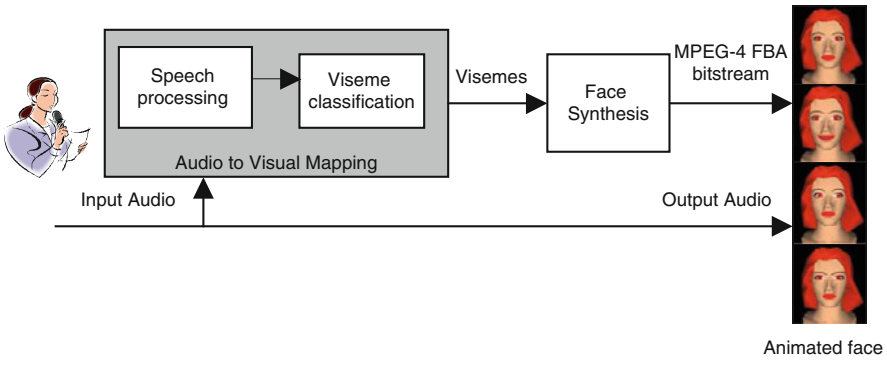

Fig. 2 Schematic view of lip sync system

2 From Emotions to Social Signals and Contexts of Interaction

Features extracted from human faces and bodies, utterances, and physiological signals may also be exploited to provide cues not necessarily related to emotion, but also in an emerging field termed "social signal processing." Here, low-level features, such as prominent points around the eyes, may be used to estimate the user's eye gaze direction, instead of going directly to a high-level concept, such as emotions and from that, the user's degree of engagement to an interacting party or machine (Vertegaal et al., 2001). This effectively demonstrates the strong relation of choosing class names for a machine learning algorithm to the underlying theoretical concepts; detected features from a human face may correspond to wide open eyes, but what *that* fact means is still a question to be answered.

Another open research question related to understanding features and adapting classifiers is the choice of context. Instead of providing "all-weather" feature extraction and recognition systems, the current (and more interesting, user-wise) trend is to exploit what knowledge is available, related to who the user is, where the interaction takes place, and in what application context, in order to fine-tune a general purpose algorithm and choose the relevant, prominent features to track (Karpouzis and Kollias, 2007). In this framework, signal processing techniques may be used in association with knowledge technology concepts to provide features for cognitive structures, effectively bridging diverse disciplines into one, user-centered loop. In addition to this, information from available modalities can also be used to fortify the result in one particular unimodal recognition; in Morency et al. (2007), authors investigate the presence of lexical hints related to posing a question (e.g., subject–verb reversal) in relation to detecting particular body gestures (a head nod), while in Christoudias et al. (2006), features from strong, successful classifications from the audio channel are used to train a visual recognition algorithm in an unsupervised manner.

References

Castellano G, Kessous L, Caridakis G (2008) Emotion recognition through multiple modalities: face, body gesture, speech. In: Peter C, Beale R (eds) Affect and emotion in human–computer interaction. Springer, Heidelberg, pp 92–103

Chen T, Rao R (1998 May) Audio-visual integration in multimodal communication. Proc IEEE Spec Issue Multimedia Signal Process 86:837–852

Christoudias M, Saenko K, Morency L-P, Darrell T (2006) Co-adaptation of audio-visual speech and gesture classifiers. In: Proceedings of the 8th international conference on multimodal interfaces, Banff, AB, Canada

Gunes H, Piccardi M (2009) Automatic temporal segment detection and affect recognition from face and body display. IEEE Trans Syst Man Cybern B 39(1):64–84

Karpouzis K, Kollias S (2007) Multimodality, universals, natural interaction, Humaine plenary presentation. http://tinyurl.com/humaine-context. Accessed 6 Jul 2009

Lewis JP, Parke FI (1986) Automated lip-synch and speech synthesis for character animation. SIGCHI Bull 17(May):143–147. doi: 10.1145/30851.30874 http://doi.acm.org/10.1145/30851.30874

McAllister DF, Rodman RD, Bitzer DL, Freeman AS (1997) Lip synchronization for animation. In: Proceedings of SIGGRAPH 97, Los Angeles, CA

Morency L, Sidner C, Lee C, Darrell T (2007) Head gestures for perceptual interfaces: the role of context in improving recognition. Artif Intell 171(8–9):568–585

Onat S, Libertus K, Koenig P (2007) Integrating audiovisual information for the control of overt attention. J Vis 7(10):1–16

Oviatt S (1999) Ten myths of multimodal interaction. Commun ACM 42(11):74–81

Pandzic IS, Forchheimer R (eds) (2002) MPEG-4 facial animation – the standard, implementation and applications. Wiley, New York, NY. ISBN 0-470-84465-5

Rogozan A (1999) Discriminative learning of visual data for audiovisual speech recognition. Int J Artif Intell Tools 8:43–52

Teissier P, Robert-Ribes J, Schwartz JL (1999) Comparing models for audiovisual fusion in a noisy-vowel recognition task. IEEE Trans Speech Audio Process 7:629–642

Vertegaal R, Slagter R, van der Veer G, Nijholt A (2001) Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In: Proceedings of the SIGCHI conference on human factors in computing systems, Seattle, Washington, DC, pp 301–308

Vinciarelli A, Pantic M, Bourlard H (2009 November) Social signal processing: survey of an emerging domain. Image Vis Comput 27(12):1743–1759

Zoric G (2003) Real-time animation driven by human voice. In: Proceedings of ConTEL, Zagreb

Zoric G, Pandzic I (2006) Real-time language independent lip synchronization method using a genetic algorithm. Signal Process 86(12):3644–3656